BIG DATA COMPUTING

ID's last digit: 0 – 4

Andrea Pietracaprina

Department of Information Engineering

University of Padova

andrea.pietracaprina@unipd.it

OUTLINE

1. Big Data Phenomenon
2. Computational Challenges
3. Organization of the Course
4. Administrative Issues

Big Data Phenomenon

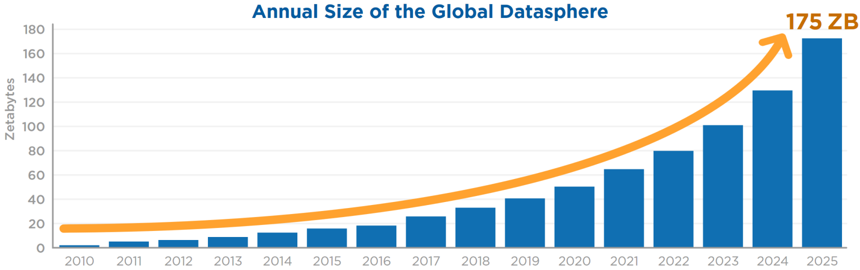
"Space is big. Really big"

(*Douglas Adams, The Hitchhiker's Guide to the Galaxy*)

Why is DATA growing so much?

- Technological progress:
  - Growth of storage capacity
  - Growth of comunication bandwith
  - Growth of computing capacity
- Reduction of ICT costs
- Pervasiveness of digital technologies: scientific research, health, business, politics, social interactions, ...

Big Data Phenomenon



From: *The Digitization of the World (IDC, 2018)*

How big is 175ZB?:

- 1 ZettaByte (ZB) = 1 trillion GB = 10^{12} GB;
- 175 ZB $\equiv$ 23 parallel stacks of DVD from Earth to Moon;
- Downloading 175 ZB at 1Gb/s takes $>$ 43 million years

Big Data Phenomenon

The world continuously collects huge amounts of:

- Physical data: from sensors, telescopes, particle physics experiments.
- Biological/medical data: from genetic studies, patient monitoring, epidemic evolution analyses.
- Human activity data: from social networks, mobile devices, internet/web traffic, IoT systems.
- Business data: from online stores, customer profiling, bank/credit-card/financial services, quality-of-service monitoring.

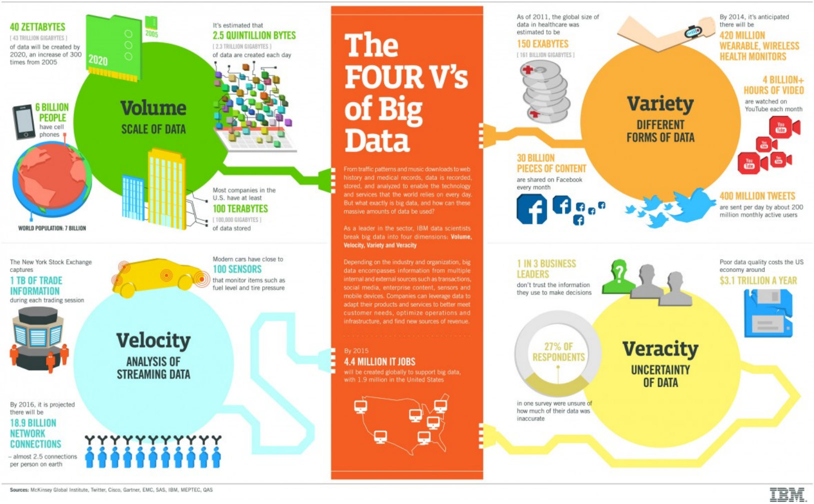
Big Data Phenomenon

The term Big Data relates to two distinct issues:

- ISSUE 1:
  - Data produced everywhere;
  - Need for automated analytics (vs human inspection);
  - Challenges: identification of suitable analysis tools, data selection/preparation.
- ISSUE 2:
  - Massive datasets need to be processed;
  - Traditional (algorithmic) approaches are unsuited;
  - Challenges: development of novel computing frameworks, novel solutions

This course focuses on ISSUE 2!

Computing Challenges



Source: IBM Big Data & Analytics Hub

Computational Challenges

- **Volume:** processing huge datasets poses several challenges and requires a data-centric perspective.
- **Veracity:** large datasets coming from real-world applications are likely to contain *noisy, uncertain data*, hence accuracy of solutions must be reconsidered.
- **Velocity:** sometimes, the data arrive at such a high rate that they cannot be stored and processed offline. Hence stream processing is needed.
- **Variety:** large datasets arise in *very different scenarios*. More effective processing is achieved by adapting to the actual characteristics of data.

The above issues require a

paradigm shift w.r.t. traditional computing.

Computational Challenges

To tackle the above challenges effectively, one needs:

- Platforms with:
  - High storage capacity and computing power
    $\Rightarrow$ parallel/distributed architectures
  - Moderate costs
  - Ease of programming and management
- Focus on accuracy-resource tradeoffs, to cope with size, noise, and uncertainty of data
- Data-centric view
- Data stream processing (sometimes)

Big Data Computing Course

What will we learn?

1. Novel computing/programming frameworks for big data processing: theory and practice
2. Key techniques to process large-scale data
   - Rigorous setting (provable guarantees)
   - Application to fundamental data analysis primitives

Specific topics

1. Distributed Frameworks: MapReduce, Apache Spark (*partitioning; resource-accuracy tradeoffs*).
2. Reducing input size: clustering (*coreset approach*, *sampling*).
3. Reducing output size: frequent itemsets (*redundancy reduction*, *top-k approach*).
4. Streaming Framework: moments estimation (*reservoir sampling*, *probabilistic counting*, *sketches*).

Organization of the Course

Subdivision into classes

The students are subdivided into two parallel classes based on their ID's last digit (*same syllabus, homeworks, and exams*)

- Class 1 (prof. Pietracaprina): last digit 0-4
- Class 2 (prof. Silvestri): last digit 5-9

Lectures

Lectures will be online via Zoom. For Class 1 meeting ID is

https://unipd.zoom.us/j/82862106641

- Lectures will be recorded and uploaded to Moodle.
- For each topic, partial slide sets are made available in advance. Final versions (together with solutions to exercises) are uploaded after the topic is fully covered.
- *Attendance and active participation are strongly encouraged*.

Organization of the Course

Exam

- Final written exam (25 points)
- Homeworks: programming assignments (7+1 points)
  - Groups of 2-3 students (even from different classes)
  - 3 homeworks, approx. one every 3 weeks.
  - Use of Apache Spark on individual PCs (Homeworks 1-2) and on CloudVeneto (Homework 3)
  - All group members receive the same grade. The extra point is given if all homeworks submitted by the respective deadlines.

Organization of the Course

Required background

- Java (preferred) or Python programming
- Basic algorithmics: asymptotic, worst-case analysis; sorting; fundamental algorithms and data structures; (e.g., lists, queues, stacks, hash tables, maps/dictionaries)
- Basic math tools, combinatorics, and probability.

Administrative Issues

Online tools

- **Course Moodle:**

  https://elearning.dei.unipd.it/course/view.php?id=6997

  - Announcements and student forum.
  - Infos: Zoom, contacts, textbooks, exam rules and sessions.
  - Material: slides, videos, exercises, articles.
  - Preliminary exams grades.
- **Uniweb:** Official exam lists and final grades.
- **Exam Moodle (only one for the two classes):**

  https://esami.elearning.unipd.it/course/view.php?id=2242

  - Submission of Homeworks.
  - Written tests (if online).

Administrative Issues

Contacts and office hours

- Teacher (prof. Andrea Pietracaprina):
  andrea.pietracaprina@unipd.it
- TAs (dott. Diego Santoro and dott. Ilie Sarpe):
  bdc-course@dei.unipd.it

Office hours (via Zoom) are by appointment (via Email). Teaching assistants should be contacted only for questions related to homeworks.

TODO: As soon as possible

- Each student must register in the Course Moodle (no password)
- Each student must register in the Exam Moodle (no password)
- Students must form groups of size at most 3 for the homeworks. Once a group is formed, it must be registered in the Exam Moodle using the Group registration link.